ICTNET at Context Suggestion Track TREC 2014

Zihao Lu[123], Zhijie Qiu[123], Liang Chang[123],Bingyang Liu[123], Dayong Wu[12], Yue Liu[12] , Xueqi Cheng[12]

1)Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190

2)Key Laboratory of Web Data Science and Technology,CAS,100190

3) University of Chinese Academy of Sciences , Beijing, 100190

{luzihao,qiuzhijie,changliang,liubingyang }@software.ict.ac.cn; {qudayong,liuyue,cxq}@ict.ac.cn

This year we did not use ClueWeb12 or ClueWeb12-B,while we solve this issue based on data we crawled from openweb.Firstly, we use external structured resource –Google Place API[1] to find all of the possible candidate places in the distance of 5 hours' drive. Secondly, we use Yandex[2] to find a description of each place because we get URL of the corresponding place. Then, we classifythe descriptions of all places. Finally, we ranked the pages based on users' preferences.

1 Data Preparation

Google Place API is used to generate candidate places for each query. We get pages from Google Place API and get descriptionsof them from Yandex. It is a challengeto get enough candidate places due to the visiting limitation of API. Another challenge is that Google Place API does not return enough candidate places for every specific geographical position. So the places we got from the API of current round are used as seeds to get more nearby places in the next round. After several rounds, we remove the duplicate places and get the candidate set.

2 Query Generation

Queries are the basics of our scheme. We want to build good queries for the search engine to return us with good suggestions.Each query consists of base part and keyword part.The base part is geographical information providing the longitude and latitude. In the base part we use Google Places to generate basic candidates. Then with the help of Yandex we can almost get the brief description of each item. By runningprocedure "lda" we generate a query vector based on LDA model. Each candidate description will generate a topic vector. By runningprocedure "cat", the query is generated by mapping every example into a vector of 10 topics, which is done manually. Then for each user an average vector of all the examples rated 5 or 4 is generated.

3 SchemeI

Procedure(RUN) "cat" consists of two steps. The first step is generating users topic vectors manually. At first we map each example to a 10-dimension vector, then a user's profile vector is defined as:

$$P_j = \frac{\sum_{i=1}^{n} v_{i,j}}{n} (1 \le j \le 10)$$

P_j is the j-th dimension of user's profile vector. n is the number of examples rated 4 or 5 by the user.

In this method we chose a ten dimension feature vector to represent a place. Every dimension is a type information that we think it's important when we do suggestions. Then we got type information for every candidate place from Google Place API. But these type tags from API are a little different from our dimensions of feature vector. So we convert these class tags into our feature vector by define the weight that map the class tag to every dimension of feature vector. We calculate each dimension of a candidate place by the following equation:

$$I_j = \frac{\sum_{i=1}^{m} v_{i,j}}{m} (1 \le j \le 10)$$

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
NOV 2014		**00-00-2014 to 00-00-2014**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
ICTNET at Context Suggestion Track TREC 2014	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Chinese Academy of Sciences,Institute of Computing Technology,Beijing 100190,**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
presented in the proceedings of the Twenty-Third Text REtrieval Conference (TREC 2014) held in Gaithersburg, Maryland, November 19-21, 2014. The conference was co-sponsored by the National Institute of Standards and Technology (NIST) and the Defense Advanced Research Projects Agency (DARPA).

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**3**	

$I_j$is the j-th dimension of the item's profile vector. m is the number of sub class which is labeled on the item.Then with the user's profile vector and the item's topic vector, we can compute a score of how much does an user like a candidate item:

$$Score_{u,item} = \sum_{i=1}^{catenum} P_i^u \cdot I_i^{item}$$

$Score_{u,item}$ is the score of how u(user) like item(item). $P_i^u$is the i'th dimension of user's profile vector P, I_i^{item} is the i'th dimension of item vector.

For a user, we compute all the score of candidates, then we recommend the top K items to the user.

4 SchemeII

Procedure(RUN) "lda" applies Topic Model to users' interests and types of places. We use descriptions of places as the input of LDA. It means that we regard each description of a place as a document. We want to find the latent type distribution of places. Sowe use LDA to find latent sematic vector for every place to represent its text descriptions. In some degree, it looks like to reduce the dimensions of the original places' description.

After running LDA model[3], we get a vector for every place. For each query, we know the users' visited places and the candidate set. We notate the user's visited place set as V. We use $rate(u,p)$ to represent the rate given by user u to place p.Then we calculate the score of the i-th candidate place for user u in the following equation:

$$Score(u,p,i) = \sum_{p \in V} \text{rate(u, p)} * \text{cosine(vector(p), vector(i))}$$

In the formulation above, we use cosine distance to represent the similarity of two vectors. Thus we get the scores of candidate places. Then we sort these places by their scores in descending order and recommend the top 50 places to user when he is in the corresponding geographical position.

5 Summary

In our schemes, we firstly used Google Place API to generate candidate set for each specified longitude and latitude. Then we tried human annotation or Latent Topic Model to find the feature vector of each place which represents its type information. Our idea is based on the fact that the user would like to visit the places with the same type that he had ever visited. For instance, if you had visited parks for several times, then you will have much more chance to visit parks in your nearby places. So in both Scheme I and Scheme II, we try to use human annotations or Latent Topical Model to find type information of places and users' interests.

After analyzing the results of our model, we found that there were some points we need to be improve. First we should get more candidate places for each specific geographical location. It means that we should make our candidate set much bigger. Because of the access limitation of Google Place API, we didn't fetchenough candidate places nearby the specified position. Second we should model the type of places in a finer grained way. More parameters should be trained and tested for LDA model. Thus we can get much accurate latent semantic type for each place. Then the representation of place will be much more useful. The same improvement can also be made in SchemeI.

To summarize, we did context suggestion by modeling interests of user and types of places. So the more accurate the type of every place is, the better suggestions will be made.

6 Acknowledgements

We would like to thank all organizers and assessors of TREC and NIST. This work is sponsored by 973 Program of China Grants *No.2013CB329602&No.2014CB340401*, 863 program of China Grants No. *2013AA01A213&No.2014AA015103*, NSF of China Grants *No.61232010&No.61173008* , and by the National Key Technology R&D Program Grants *No.2012BAH39B04 &No.2012BAH46B04*.

References

[1] Google Place API, https://developers.google.com/places/documentation/

[2] Yandex Rich Content API, http://api.yandex.com/rca/

[3] How to using gensim to get topics of documents,http://radimrehurek.com/gensim/